(GIGA)n SCIENCE

Genome sequence of *Malania oleifera*, a tree with great value for nervonic acid production

Chao-Qun Xu[1‡], Hui Liu[1‡], Shan-Shan Zhou[1‡], Dong-Xu Zhang[2], Wei Zhao[1], Sihai Wang[3], Fu Chen[4], Yan-Qiang Sun[1], Shuai Nie[1], Kai-Hua Jia[1], Si-Qian Jiao[1], Ren-Gang Zhang[5], Quan-Zheng Yun[5], Wenbin Guan[1], Xuewen Wang[4,6], Qiong Gao[1], Jeffrey L. Bennetzen[4,6], Fatemeh Maghuly[7], Ilga Porth[8,9,10], Yves Van de Peer[11,12,13], Xiao-Ru Wang[1,14], Yongpeng Ma[15]*, Jian-Feng Mao[1]*

[1] Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, National Engineering Laboratory for Tree Breeding, School of Nature Conservation, College of Biological Sciences and Technology, Beijing Forestry University, Beijing, 100083, China.

[2] College of Life Science, Datong University, Datong, 037009, Shanxi, China.

[3] Yunnan Key Laboratory of Forest Plant Cultivation and Utilization, State Forestry Administration Key Laboratory of Yunnan Rare and Endangered Species Conservation and Propagation, Yunnan Academy of Forestry, Kunming, 650201, Yunnan, China.

[4] The Camellia Institute, Yunnan Academy of Forestry, Guangnan, 663300, Yunnan, China.

[5] Beijing Ori-Gene Science and Technology Co. Ltd, Beijing, 102206, China.

[6] Department of Genetics, University of Georgia, Athens, GA 30602, USA

[7] Plant Biotechnology Unit (PBU), Dept. Biotechnology, BOKU-VIBT, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria.

[8] Département des sciences du bois et de la forêt, 1030, Avenue de la Médecine, Université Laval, Québec (Québec) G1V 0A6, Canada.

[9] Institute for System and Integrated Biology, Pavillon Charles-Eugène-Marchand, 1030,

Avenue de la Médecine, Université Laval, Québec (Québec) G1V 0A6, Canada.

[10] Centre d'Étude de la Forêt, 1030, Avenue de la Médecine, Université Laval, Québec (Québec) G1V 0A6, Canada.

[11] Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium

[12] VIB Center for Plant Systems Biology, 9052 Ghent, Belgium

[13] Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology Genetics, University of Pretoria, Private bag X20, Pretoria 0028, South Africa

[14] Department of Ecology and Environmental Science, UPSC, Umeå University, SE-901 87 Umeå, Sweden.

[15] Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Population, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China.

ORCIDs: Jeffrey Bennetzen: 0000-0003-1762-8307; Fatemeh Maghuly; 0000-0001-5433-0070; Ilga Porth: 0000-0002-9344-6348; Xiao-Ru Wang: 0000-0002-6150-7046; Yves Van de Peer: 0000-0003-4327-3730; Jian-Feng Mao: 0000-0001-9735-8516

‡These authors contributed equally to this paper.

*Correspondence to: mayongpeng@mail.kib.ac.cn (YPM); jianfeng.mao@bjfu.edu.cn (JFM)

Abstract

Background: *Malania oleifera*, a member of the Olacaceae family, is an IUCN Red Listed tree, endemic and restricted to the Karst region of South West China. This tree's seed is valued for its high content of precious fatty acids (especially nervonic acid). However, studies on its genetic make-up, and fatty acid biogenesis are severely hampered by a lack of molecular and genetic tools.

Findings: We generated 51 Gigabases (Gb) and 135 Gb of raw DNA sequences, using PacBio Single-Molecule Real-Time (SMRT) and 10x Genomics sequencing, respectively. A final genome assembly, with a scaffold N50 size of 4.65 Megabases (Mb) and a total length of 1.51 Gb, was obtained by primary assembly based on PacBio long reads plus scaffolding with 10x Genomics reads. Identified repeats constituted ~82% of the genome, and 24,064 protein-coding genes were predicted with high support. The genome has low heterozygosity and shows no evidence for recent whole genome duplication. Metabolic pathway genes relating to the accumulation of long chain fatty acid were identified and studied in detail.

Conclusions: Here, we provide the first genome assembly and gene annotation for *M. oleifera*. The availability of these resources will be of great importance for conservation biology, and for the functional genomics of nervonic acid biosynthesis.

Keywords: *de novo* genome assembly, vulnerable plant, *Malania*, nervonic acid, transcriptomes

DATA DESCRIPTION

Background information

Malania oleifera Chun & SK Lee (NCBI:txid397392), a 10-20 m high tree (**Fig. 1a-d**), is from the monotypic genus *Malania* of the Olacaceae family [1]. This tree is endemic to a restricted area within the Karst topography of southwest Guangxi and southeast Yunnan provinces, China. The recorded distribution range is bounded by N23°23′ - N24°28′ in latitude and E105°30′ - E107°30′ in longitude (**Fig. 1e**). This tree is called “garlic-fruit tree” or “suantouguo” (蒜头果) by local communities, due to its garlic shaped fruits. As an endemic tree and because of its natural populations being much reduced because of ongoing logging and habitat clearance, this tree species has been listed in the IUCN Red List as

"Vulnerable B1+2c" (extent of occurrence estimated to be < 20,000 km^2 and a continuing decline, observed, projected, or inferred, in numbers of mature individuals) [2], and has been assigned as a plant species with an extremely small population size (PSESP) for urgent conservation action [3]. Different mechanisms that could explain why *M. oleifera* became a vulnerable species have been proposed, such as niche specialization [4], limited germination and regeneration [5, 6], or pollination/mating system [7], as well as the biology of its pathogens [8]. However, until now, apart from a recent chloroplast genome sequence [9], only a few molecular genetic resources are available for *M. oleifera* to investigate its population structure and genetic makeup.

Besides conservation urgency, *M. oleifera* is also notable for its substantial phytochemical and phytopharmaceutical value: its seed has very high (64.5%) oil content [10, 11], and the highest-known proportion (55.70-67%) of nervonic acid ($C_{24}H_{46}O_2$, PubChem CID: 5281120). Nervonic acid is an important component in myelin biosynthesis in the central and peripheral nervous system. Myelin is generally localized to the sphingomyelin of animal cell membranes [12], where it has been proposed to enhance human brain function. Treatment of myelin disorders may attenuate or prevent various psychotic disorders [13, 14]. *M. oleifera* produces essential oils with benzyl alcohol (58.42%) and benzaldehyde (29.66%) as the main constituents as well as benzoic acid (1.49%) [10]. *M. oleifera* seeds also produce the glycoprotein *malania* which has high cytotoxic activity towards tumor cells and is one of the most potent toxins of plant origin [15]. Yet, little is known about the molecular mechanisms underlying the metabolic biosynthesis processes of these promising compounds in *M. oleifera*.

Here, we present a high-quality genome assembly for *M. oleifera*, combining PacBio single molecule long-reads and 10x Genomics linked reads. The assembled genome, its structural

and functional annotation and in-depth characterization will provide valuable tools for the genomic dissection of the species' genetic diversity and its population demography for future conservation purposes, as well as for in-depth molecular knowledge regarding biosynthesis and regulation of metabolism to promote the efficient and sustainable exploitation of this precious biological resource.

Plant material

One mature and healthy tree with abundant fruit (**Fig. 1 a, b, c, d**) was chosen as a tissue source for whole genome sequencing. The selected tree measured ~18 m in height, ~35 cm in diameter (at breast height) and is believed to be ~50 years old. This tree is located within a natural stand close to Diji Village, Jiumoxiang, Guangnan County, Yunnan Province, China (N23.90^{o} latitude, E104.90^{o} longitude, 1,402 m elevation) (**Fig. 1e**). The stand, from which the samples were taken, experienced little anthropogenic intervention and consists of trees of the same species but of different ages. Fresh leaves were sampled in September of 2017.

For RNA sequencing, leaves, fruits and seeds were sampled from healthy, high-yielding, mature trees from Funing County, Yunnan province and Leye County, Guangxi province, China, in different seasons during the years 2013-2016 (**Fig. 1e** and **Table S1**). Samples were immediately flash frozen in liquid nitrogen upon collection and transported on dry ice to Beijing Forestry University (BFU) for sequence analysis.

All samples were collected with permission from and under the supervision of local forestry bureaus. See **Table S1** and **Fig. 1** for more details.

PacBio SMRT sequencing

High-quality and high-molecular-weight genomic DNA was extracted from leaves of the selected tree, following the "~20 kb SMRTbell™ Libraries" protocol [16]. DNA was purified using the Mobio PowerClean® Pro DNA Clean-Up Kit, and its quality was assessed by standard agarose gel electrophoresis and Thermo Fisher Scientific Qubit Fluorometry. Genomic DNA was sheared to a size range of 15-50 Kb using either AMPure beads (Beckman Coulte) or g-TUBE (Covaris), and enzymatically repaired and converted into SMRTbell template libraries according to Pacific Biosciences instructions. Following this procedure, hairpin adapters were ligated after exonuclease-based digestion (of the remaining damaged DNA fragments and those fragments without adapters at both ends). The resulting SMRTbell templates were subsequently size-selected by Blue Pippin electrophoresis (Sage Sciences). Templates ranging from 15 to 50 Kb were sequenced on a PacBio Sequel instrument using S/P2-C2 sequencing chemistry (10 SMRT cells). A total of 5,778,035 PacBio long reads were generated, yielding a total of 51.15 Gb (roughly 30x coverage of the assembled genome) of single-molecule sequencing data with an average read length of 8,852 bp (**Fig. S1** and **Table S1**).

10x Genomics library preparation and Illumina sequencing

Purified high-molecular-weight genomic DNA of high quality was incubated with Proteinase K and RNaseA for 30 min at 25 °C. DNA was further purified, indexed and partitioned into barcoded libraries that were prepared using the GemCode kit (10x Genomics, Pleasanton, CA). Following the GemCode procedure, 1.0 ng of DNA was used for GEM (Gel Beads in Emulsion) reactions in which DNA fragments were partitioned

into molecular reactors to extend the DNA and to introduce specific 14-bp partition barcodes. Subsequently, GEM reactions were PCR-amplified. The PCR cycling protocol was: 95 °C for 5 min; cycled 18x: 4 °C for 30 s, 45 °C for 1 s, 70 °C for 20 s, and 98 °C for 30 s; held at 4 °C. The PCR products were purified as described in the GemCode protocol. Purified DNA was sheared, end-repaired, adenylation tailed, universal adapter ligated and samples indexed according to the manufacturer's recommendations.

The whole genome GemCode library was sequenced using 2x150 paired-end (PE) sequencing on Illumina HiSeq X Ten. A total of 899.778 million reads (~134.97 Gb, roughly 89x coverage of the assembled genome) were obtained, of which 89.1% had base quality values over 20 and 80% over 30 (**Table S1**). There were 19,319,151 (99.98% of total read pairs) indexes assigned to more than one read pair, while 27,368 (9.55%), 830 (2.12%) and 450 (1.80%) had more than 1000, 3000, or 5000 read pairs, respectively (**Table S2**). Sequence data were analyzed using the GemCode Long Ranger Software Suite [17, 18].

RNA sequencing

Frozen tissues were ground with a mortar and pestle, and RNA was isolated using the NEBNext Poly (A) mRNA Magnetic Isolation Module. RNA quality was determined on an Agilent 2100 BioAnalyzer. Seven sequencing libraries were prepared using the NEBNext Ultra RNA Library Prep Kit for Illumina. 150/100 bp PE sequencing was performed on an Illumina HiSeq 2000/2500 machine. See **Table S1** for details.

Estimation of genome size, heterozygosity, and repeat content

Canu v1.6 (Canu, RRID:SCR_015880) [19] was employed to filter and correct the PacBio reads. Next, k-mers were counted using Jellyfish (Jellyfish, RRID:SCR_005491) [20]. Finally, gce v1.0.0 [21] was used to estimate genome size, repeat content and the level of heterozygosity. A total of 29,971,959,192 k-mers (size 17) were identified, and the peak k-mer depth obtained was 21 (**Fig. S2**). The genome size was estimated to be ~1.50 Gb (**Table S3**). The final cleaned data corresponded to about 21-fold coverage. Repeat and error frequencies were estimated to be 54.61% and 0.34%, respectively. Heterozygosity was very low (~0.06%). See **Supplementary File 1** for commands and settings.

***De novo* genome assembly and quality control**

First, primary assemblies (eight from PacBio long reads, one from 10x Genomics linked reads) were prepared by different pipelines. Next, scaffolding and polishing were performed on the optimal primary assemblies in order to obtain a final genome assembly. Primary assembly v0.1 was generated from PacBio long reads after correction by Canu v1.6 [19], assembly v0.2 by MECAT v1.1 [22], assembly v0.3 by miniasm v0.2-r168 [23] after alignment by minimap v0.2-r124 [23], assembly v0.4 by Falcon v0.7 (Falcon, RRID: SCR_016089) [24, 25] after correction with Canu v1.6, assembly v0.5 by SMARTdenovo v1.0.0 [26] after correction with Canu v1.6, assembly v0.6 by Wtdbg v1.2.8 [27] after correction with Canu v1.6, assembly v0.7 by SMARTdenovo v1.0.0 after correction, and assembly v0.8 by Wtdbg v1.2.8. Assembly v0.9 was prepared by Supernova™ assembler 2.0 [28, 29] from 10x Genomics linked reads data. Based on quality control parameters, assembly v0.7 was chosen as optimal for further scaffolding and polishing. It generated a reasonably-sized assembly (1.51 Gb), providing the highest N50 (i.e. the shortest sequence length at 50%

of the total genome assembly length) (1.12 Mb), and the lowest number of contigs (3,038) and L50 (i.e. the smallest number of contig sequences whose lengths sum produces the N50 value) (330). Furthermore, genome assembly version v0.7 exhibited the longest contig length (6.72 Mb), as well as 71.80% gene completeness as determined by BUSCO (BUSCO, RRID:SCR_015008) [30] assessment (**Table S4**). This assembly (v0.7) was further polished with raw PacBio long reads using arrow v2.2.1 [31] to produce (in two rounds) assembly v1.0. Subsequently, 10x Genomics linked reads were processed with Long Ranger [17, 18], and were then aligned to v1.0 using BWA mem v0.7.15 (default values, *-t*12) (BWA, RRID:SCR_010910) [32] and subsequently scaffolded by ARCS v1.0.1 [33] to produce assembly v1.1. The final assembly was generated after one further iteration of polishing with arrow v2.21 and three iterations with Pilon v1.22 (Pilon, RRID:SCR_014731) [34]. Before arrow-based polishing, PacBio raw reads were aligned using BLASR v5.1 (BLASR, RRID:SCR_000764) [35, 36], and PacBio raw reads were mapped with Bowtie2 v2.2.6 (Bowtie2, RRID:SCR_016368) [37] before each iteration with Pilon. In the final assembly, a genome size of 1.51 Gb was obtained, consisting of 2,987 contigs, 1,277 scaffolds (with contig N50 of 1.22 Mb, scaffold N50 of 4.65 Mb, longest contig of 6.7 Mb and longest scaffold of 25.1 Mb), and has a gene completeness of 90.60% (**Table 1** and **Table S4**).

The consistency of the predicted genome size based on k-mer characterization and the assembled genome indicated a good quality for our assembly. Furthermore, when all clean Illumina reads were mapped to the final assembly (v1.2f), a high sequence coverage of 98.5% was obtained. In addition, an even higher sequence coverage of 99.32% was observed for mapping PacBio long reads to the final assembly using BLASR. These two coverage values suggested high sequence completeness and fidelity of the genome assembly. Mapping rates (91-98%) were also very high for transcriptomic datasets mapped to the final assembly, of which most (79-96%) were uniquely mapped (**Table S1**), with the exception of one RNA

sequencing library (SRA accession: SRR7221534) that yielded low mapping rates (10.31%), a result that we cannot explain by anything aside from microbial or other contamination (**Supplementary File 1** for commands and settings).

Transposable element and other repeat annotation

De novo repeat identification was pursued with RepeatModeler v1.0.10 (RepeatModeler, RRID:SCR_015027) [38], which employs two complementary computational methods (RECON v1.08 and RepeatScout v1.0.5 (RepeatScout, RRID:SCR_014653) [39]) for identifying repeat element boundaries and family relationships from sequence data. Subsequently, the outputs from RepeatModeler and the RepBase library [40] were combined and used for further characterization of transposable elements (TEs), many of which are not repetitive, and other repeats by homology-based methods, including identification with RepeatMasker (v4.0.7, rmblast-2.2.28) (RepeatMasker, RRID:SCR_012954) [41]. In sum, a high percentage of the genome (82.05%) was predicted to be TEs and/or repeats in the assembled genome, predominantly (65.45%) known TEs, with 11.94% uncharacterized TEs, and a smaller number (3.64%) of simple repeats. Long terminal repeat-retrotransposons (LTR-RTs) represented the highest proportion (58.23%) of the genome, while LINE (3.67%), SINE (0.11%), DNA (3.32%) and RC (0.12%) TEs made up a minor fraction (7.22%) of the genome. *Copia* (29.51% of the genome sequence) and *Gypsy* (28.15%) LTR-RTs were about equally abundant. Repeat annotations are provided in **Fig. 2a** and **Table S5.**

Transcriptome assembly and candidate gene annotation

In total, 313.36 million raw reads from RNA analyses were generated from leaf, seed, and fruit tissues and used for gene annotation (**Table S1**). Illumina raw reads were processed by

Trimmomatic v0.33 (Trimmomatic, RRID:SCR_011848) [42] and Cutadapt v1.13 (Cutadapt, RRID:SCR_011841) [43] and aligned to the genome assembly using HiSat2 v2.1.0 (HiSat2, RRID:SCR_015530) [44]. Base quality was assessed with FastQC (FastQC, RRID:SCR_014583) [45] before and after data cleaning. Statistics for the RNA sequencing data are shown in **Table S1**. Reference genome-guided and *de novo* transcriptome assemblies, respectively, were constructed with StringTie v1.3.3b (StringTie, RRID:SCR_016323) [46] and Trinity v2.0.6 (Trinity, RRID:SCR_013048) [47]. Then, transcriptome assemblies were combined and further refined using CD-HIT v4.6 (CD-HIT, RRID:SCR_007105) [48]. Finally, 57,299 unique transcripts were predicted. The summary of transcriptome assemblies is reported in **Table S6**.

For *ab initio* gene prediction, AUGUSTUS v3.2.3 (AUGUSTUS, RRID:SCR_008417) [49, 50] was employed, using model training based on coding sequences from *Arabidopsis thaliana* and 1,440 single copy orthologs from the BUSCO embryophyta_odb9 database. For evidence-based gene prediction, the individual transcripts from RNA sequencing as well as the transcriptome assembly were aligned to the repeat-masked reference genome assembly with BlastN (BLASTN, RRID:SCR_001598) and TblastX (TBLASTX, RRID:SCR_011823) from BLAST v2.2.28+ (NCBI BLAST, RRID:SCR_004870) [51] (E-value cutoff of 10^{-5}), respectively. Protein sequences from *A. thaliana* [52], *Vitis vinifera* [53], *Solanum lycopersicum* [54] and *Olea europaea* [55] were aligned to the TE-masked and repeat-masked reference genome assembly with BlastX (BLASTX, RRID:SCR_001653) (E-value cutoff of 10^{-5}). After optimization with Exonerate v2.4.0 (Exonerate, RRID:SCR_016088) [56, 57], gene model predictions were finalized using the MAKER package v2.31.9 (MAKER, RRID:SCR_005309) [58] within AUGUSTUS. AED (Annotation Edit Distance) scores were calculated for each of the predicted genes as part of the MAKER pipeline to assess the quality

of gene prediction. Putative functions for each identified gene were predicted by homology searches with BLAT (BLAT, RRID:SCR_011919) [59] against the UniProt database (UniProt, RRID:SCR_002380) [60]. Protein annotation against Pfam (Pfam, RRID:SCR_004726) [61, 62] and InterProScan (InterProScan, RRID:SCR_005829) [63] were also conducted using the scripts provided in the MAKER package. The completeness of gene annotation was checked using the BUSCO dataset (i.e. the 1,440 single-copy orthologs from the embryophyta_odb9 database) with 10^{-5} as BLAST E-value cutoff (**Supplementary File 1** for commands and settings).

A total of 24,094 genes were predicted, with average lengths of gene regions, genes (including 5', 3' UTRs, exons and introns), CDS and exons, respectively, of 11,809 bp, 1,460 bp, 1,281 bp and 244 bp (**Table S7**). The distribution of AED tagged by MAKER is shown in **Fig. S3**, in which about 83.39% of the annotated genes (20,092 genes) had an AED < 0.5 (**Table S7**), indicating well-supported gene annotation. The result from BUSCO assessment of genome assembly and annotation qualities are shown in **Table S8**. Identification of 92.29% of the universal single-copy genes (1,329 genes out of the total 1,440 genes) supported the high quality of the genome assembly. Among the 1,329 BUSCO conserved single-copy genes detected in the assembled genome, 1,217 (84.51% of the completed genes) were found to be single-copy, while 41 genes (2.85%) were complete and duplicated (**Table S8**).

The predicted genes were annotated using seven functional databases: (1) the NCBI non-redundant protein database (NR) [64], (2) the Swiss-Prot protein database [60, 65], (3) the Translated EMBL-Bank (part of the International Nucleotide Sequence Database Collaboration, TrEMBL) [60, 66], (4) the protein families database (Pfam) [67], (5) the Cluster of Orthologous Groups for eukaryotic complete genomes (KOG) database [68], (6) the KO (the Kyoto Encyclopedia of Genes and Genomes, Orthology) database (KEGG, RRID:SCR_012773) [69, 70], and (7) the Gene Ontology (GO) database (GO,

RRID:SCR_002811) [71, 72]. By this combined strategy, 91.60% of all predicted genes could be annotated with the following protein related database outcomes: NR (57.20%), Swiss-Prot (90.60%), TrEMBL (91.40%), Pfam (76.80%), KOG (87.60%), KO (32.90%), and GO (78.70%) (**Table S9**).

Differential proliferation, age dynamics and gene proximity of different LTR-RT families

LTR-RTs (58.23% of the annotated genome) represent the most abundant group of TEs in the genome of *M. oleifera*. We further examined their classification, age distribution, birth and death. LTRharvest [73] and LTRdigest [74] were used for *de novo* prediction of LTR-RTs. In this workflow, it was required that a candidate LTR-RT was separated by 1 to 15 Kb from other candidates and flanked by a pair of putative LTRs, which could range from 100 to 3,000 bp, but with a similarity >80%. The LTR-RT candidates that possessed complete Gag-Pol protein sequences were retained as intact LTR-RTs (*I*), while solo-LTRs (*S*) and truncated LTRs (*T*), were identified based on sequence similarity to the intact LTR-RTs. LTR homologies were identified by BLASTN analysis [51] with an E-value cutoff of 1e-10, 90% overlap in length and 90% identity. Further, 3 Kb of sequence data both upstream and downstream of each detected LTR homology were extracted and compared with Gag-Pol protein sequences within the GyDB 2.0 database [75, 76] using TBLASTN (TBLASTN, RRID:SCR_011822). If at least 50% of any Gag-Pol sequence was covered by the flanking sequences with an identity > 30% and an E-value cutoff of 1e-8, the corresponding LTR was excluded from the solo-LTR list. The LTR homologies that lacked any Gag-Pol homology in both the upstream and downstream sequences were considered to be solo-LTRs. In addition, LTRs with Gag-Pol sequences on one side of flanking sequences were retained as truncated LTR-RTs. The timing of LTR-RT insertion was estimated based on the divergence between

the 5′ -LTR and 3′ -LTR of the same transposon [77]. In this procedure, each LTR pair was aligned using MUSCLE v3.8.31 (MUSCLE, RRID:SCR_011812) [78] with default settings. Kimura's two-parameter method [79] was employed with a mutation rate of 1.3e-8 substitutions yr^{-1} per site to calculate approximate insertion time [80]. Superfamily classifications within the *Gypsy* and *Copia* classes are provided in **Table S10**. Although the actual mode of LTR-RT activation and amplification is manifested at the family level [81], as defined by >80% sequence homology in the LTR-RTs, we focused on overall genome properties that could be more carefully assayed and compared at the LTR-RT superfamily level (>60% homology), with categories such as Tat and Reina of *Gypsy* or Tork and Oryco of *Copia*. The proliferation history of different superfamilies of *Gypsy* and *Copia* LTR-RTs are provided in **Figs. S4** and **S5**. The distances of intact LTR-RTs to adjacent genes were calculated, and the relationships of proximity to gene and insertion time of LTR-RTs was also examined. Gene proximity for different superfamilies of *Gypsy* and *Copia* LTR-RTs are provided in **Figs. S6** and **S7, and Table S11**. The relationship between gene proximity and insertion time for major LTR-RTs superfamilies are depicted in **Figs. S8** and **S9**.

To obtain further LTR-RT relationship insights, 5′ -LTR sequences of all LTR-RTs were compared against each other with BLASTN. Two LTRs were assigned to the same cluster if they mutually covered at least 70% of their lengths with an identity of at least 60% between them. This clustering was performed using Silix v1.2.9 [82]. Solo-LTRs (*S*) and truncated LTR-RTs (*T*) were also mapped to the same cluster containing 5' LTRs from the most similar intact LTR-RTs (*I*). Furthermore, ratios of solo-LTR-RTs and truncated LTR-RTs, respectively, to intact LTR-RTs (*S*:*I*; *T*:*I*) as well as their sums were assessed to study the removal rates of LTR-RTs over the past several million years. We further assessed the proportions of clusters with *S*:*I* values greater than three to evaluate LTR-RT deletions. The

abovementioned estimates remained consistent with or without shorter scaffolds, indicating that the draft genome assembly does not affect the results presented. To make an interspecific comparison, we also collected data on LTR-RT accumulation and removal rates for related plant species from a previous study [83], in which the same pipeline as ours was used for LTR-RTs analysis. Results of the interspecific comparison are provided in **Fig S10** and **Table S12**.

A few categories of LTR-RTs were highly abundant within the *M. oleifera* genome. Twenty-six annotated clades and one unclassified clade of *Gypsy* LTR-RTs, as well as 17 annotated clades of *Copia* were identified by querying the GyDB 2.0 database with full-length LTR-RTs of *M. oleifera*. Significant differences in their individual counts, average length, and genomic representation were found for superfamilies with both *Gypsy* and *Copia* classes of LTR-RTs (**Table S10**). *Del* is the most prevalent clade of *Gypsy* in the *M. oleifera* genome, representing 6.99% of the assembled genome. *Sire* and *Tork* are the two most abundant clades of *Copia*, representing 3.77% and 1.16% of the assembled genome, respectively. More considerable variation in average sequence length was observed for clades of *Gypsy* (4,848 - 11,592 bp) compared to those of *Copia* (4,823-9,473 bp). In sum, for most clades of both *Gypsy* and *Copia* LTR-TRs, few recent amplification were identified while a single peak of ancient amplification 2-10 million years ago (mya) were observed. Exceptionally, *Galadriel* and *Tat* superfamilies of *Gypsy* showed an active recent amplification less than one mya (**Fig. S4** and **S5**). We observed some LTR-RTs overlapping genes for most of the subgroups of *Gypsy* and *Copia*, especially for the prevalent clades: about 1,500 from the *Del* clade of *Gypsy* were found to overlap with genes; > 200 from Galadriel overlapped, and also hundreds from *Sire*, *Tork*, *Oryco* and *Retrofit* of *Copia* overlapped (**Fig. S6, Fig. S7** and **Table S11**). Except for the ones overlapping with genes, LTR-RTs were mostly distributed in regions characterized by 3-5 Kb distance to genes. In

addition, we found that gene-overlapping LTR-RTs had been generated over an extended period of time, as revealed by the insertion dates for the most representative sub-groups of *Gypsy* (**Fig. S8**) and *Copia* (**Fig. S9**).

When comparing *M. oleifera* to other related plant species with respect to LTR-RTs accumulation and removal rates, we found that the *M. oleifera* genome is characterized by the largest numbers of intact, solo- and truncated LTR-RTs. Moreover, the *M. oleifera* genome has experienced relatively low removal rates (S:I = 2.28, (S+T)/I = 2.61) as evidenced by the lowest proportion of LTR clusters with S:I > 3 (**Fig. S10** and **Table S12**). Target site duplications (TSDs), usually 5 bp of identical sequence for LTR-RTs, are the direct repeats that occur at the insertion sites of most TEs. TSDs were detected for all (24,660) intact LTR-RTs. However, they were found for only 510 (<0.1% of 56,170) solo-LTRs, indicating that these elements called "solo-LTRs" in our analysis are mostly truncated LTR-RT rather than the products of unequal homologous recombination. As expected, very few (251 out of 8,196, or about 0.3%) of the truncated LTR-RTs had TSDs. Regardless of whether an LTR-RT has been converted into a solo-LTR or a truncated LTR-RT, this still represents decay of a formerly intact LTR-RT into a non-functional (i.e., immobile) status that will eventually be fully removed by the deletions associated with illegitimate recombination [80]. Given the abundance of LTR-RTs and their proximity to genes, it will be interesting to further explore their potential influence on genome evolution and gene expression.

Orthologous genes, whole genome duplication and phylogenetic inference

OrthoMCL v2.0.9 (Ortholog Groups of Protein Sequences, RRID:SCR_007839) [84] was used to identify orthologous and paralogous gene clusters in the assembled genomes of *M. oleifera* and 14 related plant species (**Table S13**), namely *Arabidopsis thaliana* [85],

Theobroma cacao [86], *Citrus grandis* [87], *Populus trichocarpa* [88], *Eucalyptus grandis* [89, 90], *Glycine max* [91], *Vitis vinifera* [92, 93], *Solanum lycopersicum* [54], *Coffea canephora* [94], *Helianthus annuus* [95], *Beta vugaris* [96], *Nelumbo nucifera* [97], *Aquilegia coerulea* [98] and *Oryza sativa* [99]. Recommended settings were used for all-against-all BLASTP comparisons (Blast+ v2.3.056) [51] and OrthoMCL analyses. OrthoMCL analyses identified 30,367 gene families (414,518 genes involved in these analyses) based on effective database sizes of all versus all BLASTP with an E-value of 10^{-5} and a Markov Chain Clustering default inflation parameter.

The amino acid sequences of 282 orthologous protein-coding single-copy genes (**Supplementary File 2)**, identified by OrthoMCL among the 15 analyzed genomes, were acquired and aligned with MUSCLE v3.8.31 [78], employing default settings (**Supplementary File 1** for commands and settings**)**. The concatenated amino acid sequences (**Supplementary File 3**) were trimmed using trimAI v1.2 (trimal -gt 0.8 -st 0.001 -cons 60) [100] and were further used for sequence evolution model selection with ModelFinder [101]. JTT+F+R5 was selected as the best model based on all employed criteria (Akaike Information Criterion AIC, corrected AIC and Bayesian Information Criterion). To construct the maximum likelihood phylogenetic tree (**Supplementary File 1** for commands and settings**)**, IQ-TREE v1.6.7 [102] was run with the selected optimal sequence evolution model (-m JTT+F+R5) and with ultrafast bootstrapping (-bb 1000) [103, 104], and employing the Shimodaira-Hasegawa-like approximate likelihood-ratio test (SH-aLRT, -alrt 1000) [105].

Phylogenetic dating (**Supplementary File 1** for commands and settings**)** was done with the MCMCTree program of PAML v4.9h [106] with the following parameters: “burnin 100000, sampfreq 200, nsample 10000”. Rice (*O. sativa*) was defined as outgroup. The dating was calibrated against the recently summated timing of divergence [107]: the

divergence of rice from other plant genomes at 113 - 128.63 Mya (refers to MRCA (most recent common ancestor), Monocotyledoneae: Acorales - [Dioscoreales + [Liliales + [Asparagales + Aracales + Poales]]], 113 - 128.63 Mya), divergence of *N. nucifera* and *A. coerulea* from other dicots at 119.6 - 128.6 Mya (refers to MRCA, Eudicotyledoneae: Ranunculales - [Vitales + Rosids + [Caryophyllales + Asterids]], 119.6 - 128.63 Mya), and divergence of *C. canephora*, *S. lycopersicum* and *H. annuus* to the lineage formed by *A. thaliana*, *V. vinifera* and other related plants at 85.8 - 128.63 Mya (refers to MRCA, Vitales - [Rosids + [Caryophyllales + Asterids]], 85.8 - 128.63 Mya; MRCA, Rosids (minus Vitales) - [Caryophyllales + Asterids], 85.8 - 128.63 Mya; MRCA, Caryophyllales - Asterids, 85.8 - 128.63 Mya). The Molecular Clock test as implemented in MEGA X [108] rejected the null hypothesis that all tips of the tree are equidistant from the root of the tree.

All branches of the reconstructed phylogenetic tree gained high support from both Shimodaira-Hasegawa-like approximate likelihood-ratio and the ultrafast bootstrapping tests with SH-aLRT > 88 % and UFBoot > 85 %, respectively (**Fig. 2b**). The phylogenetic analysis identified the closest relationship of *M. oleifera* (Santalales) to grape (*V. vinifera*, Vitales), with the divergence time between *M. oleifera* and grape estimated at ~ 88.9798 Mya with 95% confidence intervals of 37.7394 - 108.955 Mya. *N. nucifera* (Proteales) and *Aquilegia coerulea* (Ranunculales) were forming a sister clade to all other Eudicots. The phylogenetic relationship among Ranunculales, Proteales, Santalales and Vitales is unresolved in the most recent phylogeny of the angiosperms (APG IV) [109] (http://www.mobot.org/MOBOT/Research/APweb/welcome.html, accessed at Oct. 22, 2018) [110].

Amino acid sequences of intra-specific in-paralogs constructed by OrthoMCL analyses were aligned with MUSCLE v3.8.31 [111] employing default settings. *Ks* (the number of

synonymous substitutions per synonymous site) was calculated with KaKs_Calculator v2.0 [112] under a YN model, after the conversion of protein sequence alignments into the corresponding codon alignments with PAL2NAL v14 [113]. The *Ks* distribution suggests that the *M. oleifera* genome has not undergone any recent or lineage-specific whole-genome duplication (**Fig. S11**). This finding is also supported by the low number of intra-specific collinear blocks called with MCScanX (**Fig. S12**) [114].

Of the identified OrthoMCL gene families, 6,509 gene families (194,824 genes) were shared among all of the genomes analyzed. A total of 520 gene families (2,097 genes) were found to be specific to the assembled *M. oleifera* genome when compared with the other 14 genomes (**Table S14**). Using CAFE v4.0 [95, 115], 309 gene families were detected that have expanded, while 1,528 gene families were found to have contracted in the *M. oleifera* lineage (**Fig. 2b**). Hypergeometric tests were performed to determine if specific functional categories of KEGG or GO were significantly overrepresented in the families that were significantly expanded or contracted within the *M. oleifera* genome. The expanded gene families were enriched for > 100 significant ($q < 0.05$) GO-terms of three different functional categories (Biological Process (BP), Cellular Component (CC), and Molecular Function (MF)) (**Table S15**) and seven KEGG pathways (**Table S16**). Three enriched categories were related to hormone signal transduction and to biosynthesis of tyrosine, isoquinoline alkaloid, cutin and wax, terpenoid, pantothenate and CoA, and glycine. The contracted gene families were enriched for > 400 GO-terms (**Table S17**) and 11 KEGG pathways (**Table S18**) related to various aspects of secondary metabolism, at $q < 0.05$. Results from functional enrichment analysis of rapidly evolving genes are summarized in **Table S19** (for GO enrichment) and **Table S20** (for KEGG enrichment).

Metabolic gene clusters and candidate genes for fatty acid biosynthesis pathways

It is evident that genes for numerous plant secondary metabolic pathways are sometimes densely clustered in a specific genomic region, generating biosynthetic gene clusters (BGCs) [116-118]. With the newly released and robust computational toolkit, plantiSMASH [119], 23 such BGCs related to various secondary metabolic pathways were detected (**Table S21** and **Supplementary File 4**), such as saccharide- (10 gene clusters), terpene- (4), alkaloid- (2), polyketide- (1), and lignan-polyketide (1)-related. An additional five putative BGCs were identified that could not be assigned to specific secondary metabolic pathways. The identified BGCs spanned 258 to 1,282 Kb and contained 3-8 core protein domains related to secondary metabolism.

Given the importance of fatty acid production in *M. oleifera*, we further annotated genes within the fatty acid biosynthesis pathway by querying the Plant Metabolic Network (PMN v12.5 (Plant Metabolic Network, RRID:SCR_003778) [120, 121], after enzymatic annotations for coding genes through the E2P2 package v3.1 [122]. The initial (*de novo*) fatty acid biosynthesis process mainly occurs in plastids [123] of leaf mesophyll cells, seeds, and oil-accumulating fruits in plants. In this process, acetyl and malonyl groups are condensed and further elongated to give rise to the production of 16:0-ACPs (palmitic acid) and 18:0-ACPs (stearic acid and oleic acids). After this initial process, very long chain fatty acids (VLCFAs, with 22 or more carbons) can be synthesized at the endoplasmic reticulum by sequential addition of C2 moieties from malonyl-CoA to form C18 acyl groups [124].

We detected a total of 14 genes that are predicted to function in the four reactions of the elongation cycle, including the condensation of long-chain acyl-CoA and malonyl-CoA to

form 3-oxoacyl-CoA, the reduction to 3-hydroxyacyl-CoA, the dehydration to (2E)-alkan-2-enoyl-CoA, and the final reduction to an elongated fatty acyl-CoA [124]. We detected 19 candidate genes potentially functioning in the reactions of the initial process (**Fig. S13**), and 14 genes in the subsequent VLCFA biosynthesis pathway (**Fig. 2c**). Interestingly, we found the genes of the VLCFA pathway forming two gene clusters of local duplicates, one composed of 4 genes (Maole_016461, Maole_016463, Maole_016466, and Maole_016467) and the other of two genes (Maole_017397 and Maole_017398). These six genes occurring in localized clusters are all predicted to be involved in the four key reactions of the chain elongation cycle, suggesting an important effect of local gene duplication on efficient VLCFA production. By comparison, only a few cases (one including Maole_003221.T1 and Maole_003222.T1, the other including Maole_008716.T1 and Maole_008717.T1) of localized gene duplication were found for the initial fatty acid biosynthesis pathway.

Conclusions

In sum, we provide a high quality *de novo* genome assembly, and in-depth characterization for *M. oleifera*, combining PacBio single molecule long-reads and 10x Genomics linked reads. The excellent quality of the genome assembly is supported by both the 92.29% BUSCO analysis-based single-copy gene coverage and the 99.32% (PacBio long reads), 98.5% (10x Genomics linked reads) and 91-98% (Illumina RNA sequencing reads) mapping rates of the genome and transcriptome reads. Of note, the significantly low heterozygosity of the sequenced genome was a key factor for the high continuity in genome assembly of *M. oleifera* obtained in this study. This low level of heterozygosity also suggests a high level of inbreeding in the wild population of trees that was the source of genomic DNA used for genome analysis. The novel genomic resources generated in the present study

provide vital foundation for further studies on the genetics of metabolite biogenesis, the genetic basis of the vulnerable status, the significance of local gene duplications in genomes without a recent whole genome duplication, and for biotechnology aiming at an efficient exploration of valuable plant compounds. The pattern of birth-death dynamics and gene proximity of LTR-RTs, revealed here, provide a basis for future LTR-RTs studies in plants. It will be particularly interesting to investigate whether the observed slow rate of LTR-RT amplification and removal are related to the long-lived perennial lifestyle of this largely undomesticated tree species. As the only whole genome and the second genome released for the Olacaceae family and in the Santalales order, the present data resource is also of critical value for phylogenomic and comparative genomic studies.

Availability of supporting data

The genome assembly, annotations, and other supporting data are available via the *Giga*Science database *Giga*DB[125]. The raw sequence data have been deposited in the Short Read Archive (SRA) under NCBI BioProject ID PRJNA472200. All commands and parameter settings for genome assembly, quality assessment of genome assembly, transcriptome assembly from RNA-seq, repeat and gene annotation, ortholog identification, phylogenetic reconstruction and dating been uploaded to protocols.io[126].

Abbreviations

BGCs: biosynthetic gene clusters; bp: base pair; BUSCO: benchmarking universal single-copy orthologs; CDS: coding sequence; Gb: gigabases; Kb: kilobases; LTR: long terminal repeat; Mb: megabases; mya: million years ago; PSESP: plant species with an extremely small population; RT: retrotransposons; SMRT: Single-Molecule Real-Time; TE: transposable element; VLCFAs: very long chain fatty acids.

Acknowledgments

This study was funded by Fundamental Research Funds for the Central Universities (NO. YX2013-41), by the construction of the workstation for Academician Bennetzen (NO. 2015AC018), by the Science Fund of China's Yunnan government (NO. 2015BB018), and by the State Key Laboratory of Phytochemistry and Plant Resources in West China (NO. P2015-KF11).

Author Contributions

JFM, YM and JLB conceived and designed the study; CQX, HL, SSZ, ZW, SQJ, SW, FC, YQS, SN, KHJ, DZ, RGZ, WG, QG and QZY prepared the materials, conducted the experiments and analyzed all data; JFM, CQX and YM wrote the manuscript; XW, FM, IP, YVP, JLB and XRW were involved in data interpretation and finalizing the manuscript draft. All authors read and approved the final draft.

Conflict of Interest

The authors declare that they have no competing financial interests.

References

1. Wu Z, Raven P and Hong D. Flora of China. Vol. 5 (Ulmaceae through Basellaceae). Science Press, Beijing, and Missouri Botanical Garden Press, St. Louis, 2003.
2. Sun W: *Malania oleifera*. The IUCN Red List of Threatened Species 1998: e.T32361A9701100. http://dx.doi.org/10.2305/IUCN.UK.1998.RLTS.T32361A9701100.en. Accessed 08 July 2018.
3. Ma Y, Chen G, Edward Grumbine R, Dao Z, Sun W and Guo H. Conserving plant species with extremely small populations (PSESP) in China. Biodiversity and Conservation. 2013;22(3):803-809. doi:10.1007/s10531-013-0434-3.

4. Xie WD, Chen JH, Lai JY, Shi HM, Huang KX, Liu JB, et al. Analysis on relationship between geographic distribution of *Malania oleifera* and hydro-thermal factors. Journal of Tropical & Subtropical Botany. 2009;17(4):388-394. doi:10.3969/j.issn.1005-3395.2009.4.2125.

5. Xie WD, Chen JH, Lai JY, Shi HM, Lin SF, Liu B, et al. Life-table analysis of *Malania oleifera*, a rare and endangered plant. Journal of Central South University of Forestry & Technology. 2009;29(2):73-76.

6. Wu Y, Li X and Hu Y. Reproductive biology of *Malania oeifera*. Acta Scientiarum Naturalium Universitatis Sunyatseni. 2004;43(2):81-83.

7. Lai JY, Shi HM, Pan CL, Chen SW, Ye YZ, Ming LI, et al. Pollination biology of rare and endangered species *Malania oleifera* Chun et Lee. Journal of Beijing Forestry University. 2008.

8. Xiong Y, Hong L, Li H and Li X. Bionomics of the pathogens of *Malania oleifera* seed rot. Forest Pest & Disease. 2003;22:1-4.

9. Liu SS, Hu YH, Maghuly F, Porth IM and Mao JF. The complete chloroplast genome sequence annotation for *Malania oleifera*, a critically endangered and important bioresource tree. Conservation Genetics Resources. 2018; doi:10.1007/s12686-018-1005-4.

10. Tang TF, Liu XM, Ling M, Lai F, Zhang L, Zhou YH, et al. Constituents of the essential oil and fatty acid from *Malania oleifera*. Industrial Crops and Products. 2013;43:1-5. doi:https://doi.org/10.1016/j.indcrop.2012.07.003.

11. Ma BL, Liang SF, Zhao DY, Xu AX and Zhang KJ. Study on plants containing nervonic acid. Acta Botanica Boreali-occidentalia Sinica. 2004;24(12):2362-2365.

12. Sandhir R, Khan M, Chahal A and Singh I. Localization of nervonic acid beta-oxidation in human and rodent peroxisomes: impaired oxidation in Zellweger syndrome and X-linked adrenoleukodystrophy. Journal of Lipid Research. 1998;39(11):2161-2171.

13. Oda E, Hatada K, Kimura J, Aizawa Y, Thanikachalam PV and Watanabe K. Relationships between serum unsaturated fatty acids and coronary risk factors: negative relations between nervonic acid and obesity-related risk factors. International Heart Journal. 2005;46(6):975-85.

14. Amminger GP, Schafer MR, Klier CM, Slavik JM, Holzer I, Holub M, et al. Decreased nervonic acid levels in erythrocyte membranes predict psychosis in help-seeking ultra-high-risk individuals. Molecular Psychiatry. 2012;17(12):1150-1152. doi:10.1038/mp.2011.167.

15. Yuan Y, Dai X, Wang D and Zeng X. Purification, characterization and cytotoxicity of malanin, a novel plant toxin from the seeds of *Malania oleifera*. Toxicon. 2009;54(2):121-7. doi:https://doi.org/10.1016/j.toxicon.2009.03.024.

16. Preparing *Arabidopsis* Genomic DNA for Size-Selected ~20 kb SMRTbell™ Libraries. http://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf. Accessed 20 Sept 2017.

17. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature Biotechnology. 2016;34:303. doi:10.1038/nbt.3432.

18. An open-source release of Long Ranger 2.2.0. https://github.com/10xGenomics/longranger. Accessed 01 Dec 2017.

19. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH and Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. 2017;27(5):722-736.

20. Marcais G and Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764-770. doi:10.1093/bioinformatics/btr011.

21. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. arXiv preprint arXiv:13082012. 2013.

22. Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, et al. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. Nature Methods. 2017;14:1072. doi:10.1038/nmeth.4432.

23. Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics. 2016;32(14):2103-2110. doi:10.1093/bioinformatics/btw152.

24. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods. 2016;13(12):1050-1054. doi:10.1038/nmeth.4035.

25. FALCON: experimental PacBio diploid assembler. https://github.com/PacificBiosciences/FALCON/. Accessed 01 May 2018.

26. Ultra-fast *de novo* assembler using long noisy reads. https://github.com/ruanjue/smartdenovo. Accessed 01 May 2018.

27. A fuzzy bruijn graph (FBG) approach to long noisy reads assembly. https://github.com/ruanjue/wtdbg-1.2.8. Accessed 01 May 2018.

28. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct determination of diploid genome sequences. Genome Research. 2017;27(5):757-767. doi:10.1101/gr.214874.116.

29. Pipelines for a *de novo* assembly software: Supernova. https://support.10xgenomics.com/de-novo-assembly/software/overview/latest/welcome. Accessed 01 May 2018.

30. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-3212. doi:10.1093/bioinformatics/btv351.

31. A variantCaller tool to get consensus and variant calls from mapped PacBio reads. https://github.com/PacificBiosciences/GenomicConsensus. Accessed 01 Dec 2017.

32. Li H and Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26(5):589-595. doi:10.1093/bioinformatics/btp698.

33. Yeo S, Coombe L, Warren RL, Chu J and Birol I. ARCS: scaffolding genome drafts with linked reads. Bioinformatics. 2018;34(5):725-731. doi:10.1093/bioinformatics/btx675.

34. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963. doi:10.1371/journal.pone.0112963.

35. Chaisson MJ and Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13(1):238. doi:10.1186/1471-2105-13-238.

36. A long read aligner tool for PacBio. https://github.com/PacificBiosciences/blasr. Accessed 01 May 2018.

37. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9:357. doi:10.1038/nmeth.1923.

38. RepeatModeler: a *de novo* repeat family identification and modeling package. http://www.repeatmasker.org/RepeatModeler/. Accessed 01 May 2018.

39. Price AL, Jones NC and Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005;21Suppl 1:i351-i358. doi:10.1093/bioinformatics/bti1018.

40. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015;6:11. doi:10.1186/s13100-015-0041-9.

41. A program that screens DNA sequences for interspersed repeats and low complexity DNA sequences: RepeatMasker. http://www.repeatmasker.org/. Accessed 01 May 2018.

42. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170.

43. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(1). doi:10.14806/ej.17.1.200.

44. Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods. 2015;12(4):357-360.

45. A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 01 May 2018.

46. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT and Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 2015;33(3):290-295.

47. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology. 2011;29:644. doi:10.1038/nbt.1883.

48. Fu L, Niu B, Zhu Z, Wu S and Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-3152.

49. Keller O, Kollmar M, Stanke M and Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011;27(6):757-763. doi:10.1093/bioinformatics/btr010.

50. Stanke M, Diekhans M, Baertsch R and Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 2008;24(5):637-644. doi:10.1093/bioinformatics/btn013.

51. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ and Madden TL. Domain enhanced lookup time accelerated BLAST. Biology Direct. 2012;7 1:12. doi:10.1186/1745-6150-7-12.

52. Swarbreck D, Wilks C, Lamesch P, Berardini T, Garcia-Hernandez M and Foerster H. The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Research. 2007;36:D1009-D1014.

53. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449(7161):463-467. doi:10.1038/nature06148.

54. Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485(7400):635-641.

55. Cruz F, Julca I, Gómez-Garrido J, Loska D, Marcet-Houben M, Cano E, et al. Genome sequence of the olive tree, *Olea europaea*. GigaScience. 2016;5(1):29. doi:10.1186/s13742-016-0134-5.

56. Slater GSC and Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31. doi:10.1186/1471-2105-6-31.

57. A generic tool for sequence alignment. https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate. Accessed 01 May 2018.

58. Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research. 2008;18(1):188-196. doi:10.1101/gr.6743907.

59. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Research. 2002;12(4):656-664. doi:10.1101/gr.229202.

60. Bairoch A and Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Research. 2000;28(1):45-48.

61. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, et al. The Pfam protein families database. Nucleic Acids Research. 2002;30(1):276-280.

62. Punta M, Coggill P, Eberhardt R, Mistry J, Tate J and Boursnell C. The Pfam protein families database. Nucleic Acids Research. 2011;40:D290-D301.

63. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Research. 2005;33(Web Server issue):W116-W120. doi:10.1093/nar/gki442.

64. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/. Accessed 01 Dec 2017.

65. ExPASy Bioinformatics Resources Portal. http://www.expasy.ch/sprot. Accessed 01 May 2018.

66. UniProt. http://www.ebi.ac.uk/uniprot. Accessed 01 May 2018.

67. Pfam. http://pfam.xfam.org/. Accessed 01 May 2018.

68. The KOG Browser. http://genome.jgi-psf.org/help/kogbrowser.jsf. Accessed 01 May 2018.

69. Kanehisa M and Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28 1:27-30.

70. KO (KEGG ORTHOLOGY) Database. http://www.genome.jp/kegg/ko.html. Accessed 01 Dec 2017.

71. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research. 2004;32(Database issue):D258-D261. doi:10.1093/nar/gkh036.

72. Gene Ontology Consortium. http://www.geneontology.org. Accessed 01 May 2018.

73. Ellinghaus D, Kurtz S and Willhoeft U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics. 2008;9(1):18. doi:10.1186/1471-2105-9-18.

74. Steinbiss S, Willhoeft U, Gremme G and Kurtz S. Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons. Nucleic Acids Research. 2009;37(21):7002-7013. doi:10.1093/nar/gkp759.

75. Llorens C, Futami R, Covelli L, Domínguez-Escribá L, Viu JM, Tamarit D, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research. 2011;39suppl 1:D70-D74. doi:10.1093/nar/gkq1061.

76. Lloréns C, Futami R, Bezemer D and Moya A. The Gypsy Database (GyDB) of mobile genetic elements. Nucleic Acids Research. 2008;36 suppl 1:D38-D46. doi:10.1093/nar/gkm697.

77. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y and Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nature Genetics. 1998;20(1):43-45. doi:10.1038/1695.

78. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340.

79. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution. 1980;16(2):111-120. doi:10.1007/bf01731581.

80. Ma J and Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(34):12404-12410. doi:10.1073/pnas.0403715101.

81. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics. 2007;8(12):973-982.

82. Miele V, Penel S and Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011;12(1):116. doi:10.1186/1471-2105-12-116.

83. Lyu H, He Z, Wu CI and Shi S. Convergent adaptive evolution in marginal environments: unloading transposable elements as a common strategy among mangrove genomes. New phytologist. 2018;217(1):428-438. doi:10.1111/nph.14784.

84. Li L, Stoeckert CJ and Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Research. 2003;13(9):2178-2189. doi:10.1101/gr.1224503.

85. Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S and Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. The Plant Journal : for cell and molecular biology. 2017;89(4):789-804. doi:10.1111/tpj.13415.

86. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Iii DL, Cornejo O, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biology. 2013;14(6):r53. doi:10.1186/gb-2013-14-6-r53.

87. Wang X, Xu Y, Zhang S, Cao L, Huang Y, Cheng J, et al. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. Nature Genetics. 2017;49:765. doi:10.1038/ng.3839.

88. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313(5793):1596-1604. doi:10.1126/science.1128691.

89. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of *Eucalyptus grandis*. Nature. 2014;510:356-362. doi:10.1038/nature13308.

90. Bartholome J, Mandrou E, Mabiala A, Jenkins J, Nabihoudine I, Klopp C, et al. High-resolution genetic maps of Eucalyptus improve *Eucalyptus grandis* genome assembly. New Phytologist. 2015;206(4):1283-1296. doi:10.1111/nph.13150.

91. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nature Genetics. 2014;46(7):707-713. doi:10.1038/ng.3008.

92. The French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463. doi:10.1038/nature06148.

93. Canaguier A, Grimplet J, Di Gaspero G, Scalabrin S, Duchêne E, Choisne N, et al. A new version of the grapevine reference genome assembly (12X.v2) and of its

annotation (VCost.v3). Genomics Data. 2017;14:56-62. doi:10.1016/j.gdata.2017.09.002.

94. Denoeud F, Carretero-Paulet L, Dereeper A, Droc G, Guyot R, Pietrella M, et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science. 2014;345(6201):1181-1184. doi:10.1126/science.1255274.

95. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 2017;546:148. doi:10.1038/nature22380.

96. Dohm JC, Minoche AE, Holtgrawe D, Capella-Gutierrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2013;505:546-549. doi:10.1038/nature12817.

97. Ming R, VanBuren R, Liu Y, Yang M, Han Y, Li LT, et al. Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). Genome Biology. 2013;14(5):R41. doi:10.1186/gb-2013-14-5-r41.

98. Filiault D, Ballerini E, Mandakova T, Akoz G, Derieg N, Schmutz J, et al. The *Aquilegia* genome: adaptive radiation and an extraordinarily polymorphic chromosome with a unique history. eLife. 2018; 7:e36426. doi: 10.7554/eLife.36426.

99. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Research. 2007;35(Database issue):D883-887. doi:10.1093/nar/gkl976.

100. Capella-Gutiérrez S, Silla-Martínez JM and Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-1973. doi:10.1093/bioinformatics/btp348.

101. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A and Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods. 2017;14:587. doi:10.1038/nmeth.4285.

102. Nguyen LT, Schmidt HA, von Haeseler A and Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015;32(1):268-274. doi:10.1093/molbev/msu300.

103. Minh BQ, Nguyen MAT and von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Molecular Biology and Evolution. 2013;30(5):1188-1195. doi:10.1093/molbev/mst024.

104. Hoang DT, Chernomor O, von Haeseler A, Minh BQ and Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2018;35(2):518-522. doi:10.1093/molbev/msx281.

105. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W and Gascuel O. New algorithms and methods to estimate maximum likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology. 2010;59(3):307-321. doi:10.1093/sysbio/syq010.

106. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007;24(8):1586-1591. doi:10.1093/molbev/msm088.

107. Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, et al. The timescale of early land plant evolution. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(10):E2274-E2283. doi:10.1073/pnas.1719588115.

108. Kumar S, Stecher G, Li M, Knyaz C and Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology and Evolution. 2018;35(6):1547-1549. doi:10.1093/molbev/msy096.

109. Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, et al. An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society. 2016;181(1):1-20. doi:10.1111/boj.12385.

110. Stevens PF and Davis HM. The angiosperm phylogeny website - a tool for reference and teaching in a time of change. Proceedings of the American Society for Information Science and Technology. 2005;42(1). doi:10.1002/meet.14504201249.

111. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340.

112. Wang D, Zhang Y, Zhang Z, Zhu J and Yu J. KaKs_Calculator 2.0: a Toolkit incorporating gamma-series methods and sliding window strategies. Genomics, Proteomics & Bioinformatics. 2010;8(1):77-80. doi:https://doi.org/10.1016/S1672-0229(10)60008-3.

113. Suyama M, Torrents D and Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research. 2006;34(Web Server issue):W609-W612. doi:10.1093/nar/gkl315.

114. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Research. 2012;40(7):e49. doi:10.1093/nar/gkr1293.

115. De Bie T, Cristianini N, Demuth JP and Hahn MW. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22(10):1269-1271. doi:10.1093/bioinformatics/btl097.

116. Chae L, Kim T, Nilo-Poyanco R and Rhee SY. Genomic signatures of specialized metabolism in plants. Science. 2014;344(6183):510-513.

117. Nützmann HW, Huang A and Osbourn A. Plant metabolic clusters - from genetics to genomics. New phytologist. 2016;211(3):771-789. doi:10.1111/nph.13981.

118. Nützmann HW and Osbourn A. Gene clustering in plant specialized metabolism. Current Opinion in Biotechnology. 2014;26:91-109. doi: 10.1016/j.copbio.2013.10.009.

119. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A and Medema MH. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Research. 2017;45(W1):W55-W63. doi:10.1093/nar/gkx305.

120. Schlapfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, et al. Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. Plant Physiology. 2017;173(4):2041-2059. doi:10.1104/pp.16.01942.

121. PMN: a plant metabolic pathway databases. https://www.plantcyc.org/. Accessed 01 May 2018.

122. E2P2: An enzyme annotation pipeline used to generate the species-specific metabolic databases. https://gitlab.com/rhee-lab/E2P2/tree/master. Accessed 01 Dec 2017.

123. Yasuno R, von Wettstein-Knowles P and Wada H. Identification and molecular characterization of the β-ketoacyl-[acyl carrier protein] synthase component of the *Arabidopsis* mitochondrial fatty acid synthase. Journal of Biological Chemistry. 2004;279(9):8242-8251.

124. Jakobsson A, Westerberg R and Jacobsson A. Fatty acid elongases in mammals: their regulation and roles in metabolism. Progress in Lipid Research. 2006;45(3):237-249.

125. Xu CQ; Liu H; Zhou SS; Zhang DX; Zhao W; Wang S; Chen F; Sun YQ; Nie S; Jia KH; Jiao SQ; Zhang RG; Yun QZ; Guan W; Wang X; Gao Q; Bennetzen JL; Maghuly F; Porth I; de Peer YV; Wang XR; Ma Y; Mao JF (2018): Supporting data for "Genome sequence of *Malania oleifera*, a tree with great value for nervonic acid production" GigaScience Database. http://dx.doi.org/10.5524/100549

126. Chao-Qun Xu, Hui Liu, Shan-Shan Zhou, Dong-Xu Zhang, Wei Zhao, Sihai Wang, Fu Chen, Yan-Qiang Sun, Shuai Nie, Kai-Hua Jia, Si-Qian Jiao, Ren-Gang Zhang, Quan-Zheng Yun, Wenbin Guan, Xuewen Wang, Jeffrey L. Bennetzen, Fatemeh Maghuly, Ilga Porth, Yves Van de Peer1, Xiao-Ru Wang, Yongpeng Ma, Jian-Feng Mao (2018): Malania oleifera genome assembly and annotation. protocols.io http://dx.doi.org/10.17504/protocols.io.u7nezme

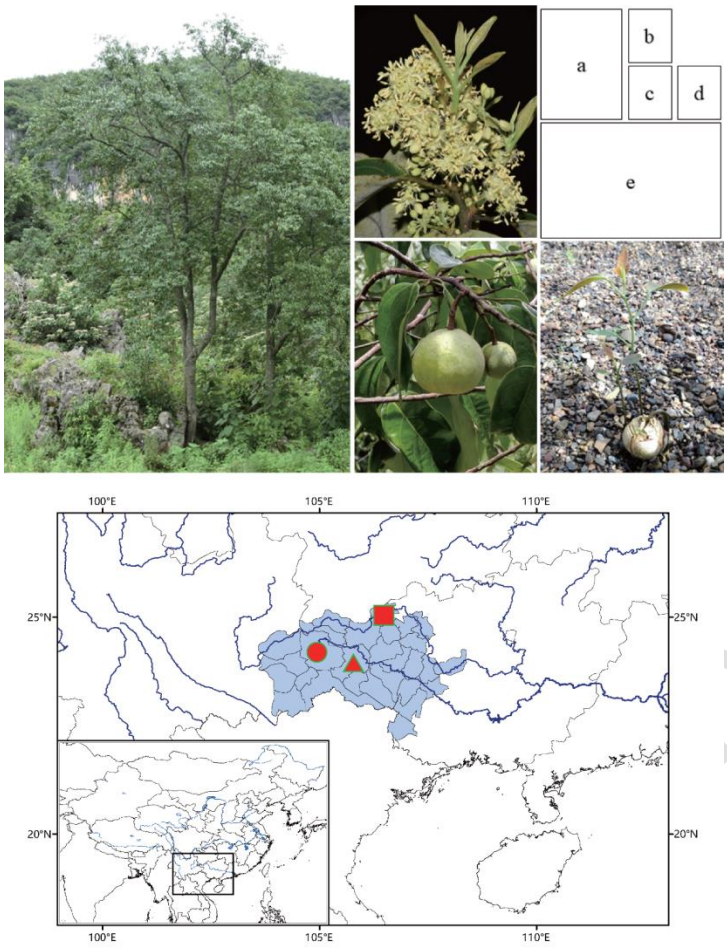

Fig. 1 Images of *M. oleifera*, recorded distribution range and sampling sites.

a-d, mature tree (a), flower (b), fruit (c) and naturally germinated seedling (d); e, blue shaded region denotes the reported distribution range of *M. oleifera*, while the red circle denotes the position (N23.90°, E104.09°, Guangnan County, Yunnan) where one tree was sampled for whole genome sequencing, and the red triangle and square denote the positions (N23.9°, E106.00°, Funing County, Yunnan and N24.78°, E106.57°, Leye County, Guangxi) where trees were sampled for RNA sequencing.

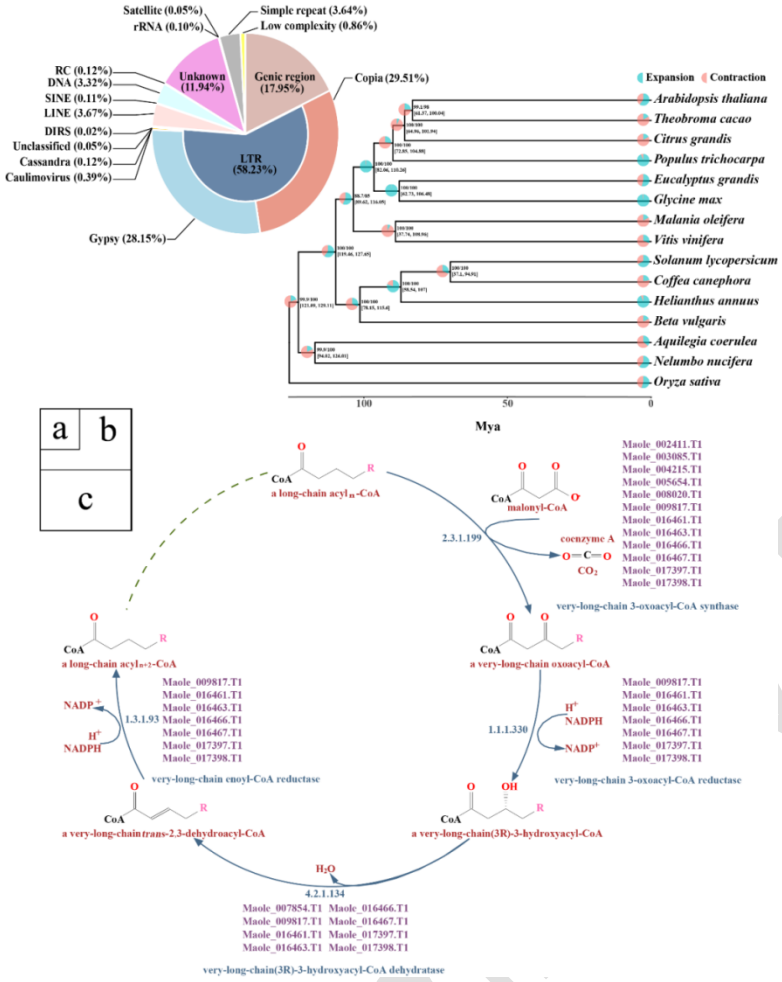

Fig. 2 Repeat composition, phylogenomic inferences and biosynthesis pathway for very long chain fatty acids synthesis in *M. oleifera*.

a. genome proportions of genic and various repeat sequences; b. phylogenetic tree, divergence time, and profiles of gene families that underwent expansion or contraction; bootstrapping supports (SH-aLRT/UFBoot) are presented along with the 95% confidence intervals for each dating point in brackets; c. annotated genes involved in the biosynthesis pathway of very long chain fatty acids (a fatty acid with minimum 22 carbon moieties) in *M. oleifera*.

Table 1. Statistics of the final genome assembly for *M. oleifera*.

	Contig		Scaffold	
	Size (bp)	**Number**	**Size (bp)**	**Number**
Total Size	1,509,344,141	-	1,519,782,615	-
Total Number	-	2,987	-	1,277
N10	2,959,726	39	11,755,999	10
N50	1,218,690	376	4,647,296	94
N90	272,293	1,337	1,153,659	339
Max.	6,703,356	-	25,060,663	-
Min.	334	-	8256	-
Mean	505,304	-	1,190,119	-
Median	200,407	-	85,436	-
Gap	-	-	10,438,474 (0.69%)	1,710
GC Content	36.07%	-	35.82 %	-

-, Data not available.

(GIGA)n SCIENCE

Supplementary Figures

Fig. S1. Length distribution of PacBio subreads.

Fig. S2. K-mer frequency distribution estimated from PacBio sequences after filtering and correction at k-mer size of 17. A k-mer refers to an artificial sequence division of K nucleotides. From k-mer frequencies, genomic characteristics (genome size, repeat structure and heterozygous rate) could be estimated. Peaks at depth of 21 are annotated with dashed lines.

Fig. S3. Distribution of AED (annotation edit distance) scores from gene prediction. AED = 0 indicates perfect agreement between the gene prediction and the transcript and protein evidence; AED = 1 indicates no evidence support for annotation.

Fig. S4. Proliferation history of different superfamilies of the *Gypsy* class of LTR-RTs (long terminal repeat-retrotransposons) in the *M. oleifera* genome.

Fig. S5. Proliferation history of different superfamilies of the *Copia* class of LTR-RTs in the *M. oleifera* genome.

Fig. S6. Gene proximity for different superfamilies of the *Gypsy* class of LTR-RTs in the *M. oleifera* genome.

The natural logarithm of the base distance between an LTR-RT and an adjacent gene (plus one) was used as the X axis.

Fig. S7. Gene proximity for different superfamilies of the *Copia* class of LTR-RTs in the *M. oleifera* genome.

The natural logarithm of the base distance between an LTR-RT and an adjacent gene (plus one) was used as the X axis.

Fig. S8. Gene proximity and insertion time for major superfamilies of the *Gypsy* class of LTR-RTs in the *M. oleifera* genome.

The natural logarithm of the base distance between an LTR-RT and an adjacent gene (plus one) was used as the Y axis, time in mya as X axis.

Fig. S9. Gene proximity and insertion time for major superfamilies of the *Copia* class of LTR-RTs in the *M. oleifera* genome.

The natural logarithm of the base distance between an LTR-RT and an adjacent gene (plus one) was used as the Y axis, time in mya as X axis.

Fig. S10. Birth and death of LTR-RTs (long terminal repeat-retrotransposons) in the *M. oleifera* genome compared to six other members of Rosids and two members from Asterids. (a) total numbers of intact LTR-RTs in the genome; (b) comparison of $S + T$ values among these nine plant species; (c) total numbers of intact LTR-RTs and traces of LTR-RT death; (d) ratios of solo-LTR to intact LTR-RT (S:I). (e) proportions of LTR-RTs found in the clusters with high removal rates (filtered S:$I \geq 3$). S, number of solo-LTRs; T, number of truncated LTR-RTs I, number of intact LTR-RTs.

Fig. S11. *Ks* distribution of paralogs in synteny blocks within the *M. oleifera* genome.

Fig. S12. Gene:synteny-block pattern in the *M. oleifera* genome.

Fig. S13. Genes annotation for the initial (*de novo*) fatty acid biosynthesis process in the *M. oleifera* genome.

(GIGA)n SCIENCE

Supplementary Tables

Table S1. Summary of PacBio and Illumina sequencing data (10x Genomics and RNA sequencing) generated in the present study. IDs of the study, sample, library and accessions in NCBI SRA and employed sequencing platform, material origins of the sequenced DNA or RNA, statistics of raw and cleaned data, and mapping rates are shown.

Table S2. Data summary from 10x Genomics analysis based on GemCode index multiplicity. Read subsets are based on the number of associated reads for each index. For raw reads, all indices (including those with N's) are included in the count. For all other read sets, only the indices without N's were used for binning.

Table S3. Estimation of genome characteristics based on 17-mer statistics.

Table S4. Statistics of the different versions of *M. oleifera* genome assembly in ascending order. N50: shortest sequence length at 50% of the genome; L50: smallest number of contigs whose length sum produces N50. NA: data not available; * statistics for contigs/scaffolds. Gene completeness was generated by assessment with 1,440 single copy orthologs from the BUSCO embryophyta_odb9 database.

Table S5. Summary of the annotated TEs in the genome assembly for *M. oleifera*. LTR: Long Terminal Repeat retrotransposons; LINE: Long Interspersed Nuclear Element, a category of non-LTR (long terminal repeat) retroelements; SINE: Short Interspersed Nuclear Element, a category of non-autonomous and non-coding retroelements (TEs); RC: Rolling-circle transposons.

Table S6. Summary of transcriptome assemblies using three different analysis pipelines.

Table S7. Summary of annotated genes.

AED: Annotation Edit Distance; gene region (including 5', 3' UTRs, exons and introns).

Table S8. Summary of BUSCO evaluation for gene prediction.

Table S9. Summary of functional annotation of predicted genes.

Table S10. Superfamilies within the *Gypsy* and *Copia* LTR-RTs classes of TEs.

Table S11. Gene proximity of superfamilies of *Gypsy* and *Copia* classes of LTR-RTs.

Table S12. Comparison of the number of original and filtered intact LTR-RT, solo-LTR and Truncated LTR TEs among 9 plant species.

Table S13. Genomic data used for phylogenomic and gene family analyses. Origins, download links, assembly versions, genome properties and references of 14 genomes are shown.

Table S14. Summary of gene family analyses. Unique groups and genes, single-copy and duplicated groups and genes are summarized for the 15 analyzed plant genomes.

Table S15. GO enrichment of expanded gene families. (A) 'Category' is the Gene Ontology (GO) term ID; (B) 'P_value' is the overrepresentation p-value indicating the observed frequency of a given term among analyzed genes is equal to the expected frequency based on the null distribution; i.e. lower p-values indicate stronger evidence for overrepresentation; (C) 'Q_value' is the Benjamini and Hochberg adjusted p-value, (D) 'numEPInCat' is the number of expanded gene families in the corresponding GO category; (E) 'numInCat' is the number of detected gene families in the corresponding GO category; (F) 'Term' is the GO term; (G) 'Ontology' indicates which ontology the term comes from. Significant at $q < 0.05$.

Table S16. KEGG enrichment of expanded gene families. (A) 'KO category' is the KEGG Orthology (KO) category ID; (B) 'P_value' is the over represented p-value indicating the

observed frequency of a given term among analyzed genes is equal to the expected frequency based on the null distribution; i.e. lower p-values indicate stronger evidence for overrepresentation; (C) 'Q_value' is the Benjamini and Hochberg adjusted p-value, (D) 'numEPInCat' is the number of expanded gene families in the corresponding KO category; (E) 'numInCat' is the number of detected gene families in the corresponding KO category; (F) 'Pathway' is the KEGG pathway; (G) 'Class' indicates which KEGG class the pathway comes from. Significant at $q < 0.05$.

Table S17. GO enrichment of contracted gene families. (A) 'Category' is the Gene Ontology (GO) term ID; (B) 'P_value' is the over represented p-value indicating the observed frequency of a given term among analyzed genes is equal to the expected frequency based on the null distribution; i.e. lower p-values indicate stronger evidence for overrepresentation; (C) 'Q_value' is the Benjamini and Hochberg adjusted p-value, (D) 'numEPInCat' is the number of expanded gene families in the corresponding GO category; (E) 'numInCat' is the number of detected gene families in the corresponding GO category; (F) 'Term' is the GO term; (G) 'Ontology' indicates which ontology the term comes from. Significant at $q < 0.05$.

Table S18. KEGG enrichment of contracted gene families. (A) 'KO category' is the KEGG Orthology (KO) category ID; (B) 'P_value' is the over represented p-value indicating the observed frequency of a given term among analyzed genes is equal to the expected frequency based on the null distribution; i.e. lower p-values indicate stronger evidence for overrepresentation; (C) 'Q_value' is the Benjamini and Hochberg adjusted p-value, (D) 'numEPInCat' is the number of expanded gene families in the corresponding KO category; (E) 'numInCat' is the number of detected gene families in the corresponding KO category; (F) 'Pathway' is the KEGG pathway; (G) 'Class' indicates which KEGG class the pathway comes from. Significant at $q < 0.05$.

Table S19. GO enrichment of fast evolving gene families. (A) 'Category' is the Gene Ontology (GO) term ID; (B) 'P_value' is the over represented p-value indicating the observed frequency of a given term among analyzed genes is equal to the expected frequency based on the null distribution; i.e. lower p-values indicate stronger evidence for overrepresentation; (C) 'Q_value' is the Benjamini and Hochberg adjusted p-value, (D) 'numEPInCat' is the number of expanded gene families in the corresponding GO category; (E) 'numInCat' is the number of detected gene families in the corresponding GO category; (F) 'Term' is the GO term; (G) 'Ontology' indicates which ontology the term comes from. Significant at $q < 0.05$.

Table S20. KEGG enrichment of fast evolving gene families. (A) 'KO category' is the KEGG Orthology (KO) category ID; (B) 'P_value' is the over represented p-value indicating the observed frequency of a given term among analyzed genes is equal to the expected frequency based on the null distribution; i.e. lower p-values indicate stronger evidence for overrepresentation; (C) 'Q_value' is the Benjamini and Hochberg adjusted p-value, (D) 'numEPInCat' is the number of expanded gene families in the corresponding KO category; (E) 'numInCat' is the number of detected gene families in the corresponding KO category; (F) 'Pathway' is the KEGG pathway; (G) 'Class' indicates which KEGG class the pathway comes from. Significant at $q < 0.05$.

Table S21. Summary of 23 metabolic gene clusters in the *M. oleifera* genome. Genomic coordinates, gene composition, core protein domains related to metabolism and pathway assignments are shown.

(GIGA)n SCIENCE

Supplementary Files

Supplementary File 1. The commands and parameter settings for all steps in genome assembly, quality assessment of the genome assembly, transcriptome assembly from RNA-seq data, repeat and gene annotation, ortholog identification and phylogenetic reconstruction and dating.

Supplementary File 2. Gene names/codes for the 282 orthologous protein-encoding single-copy genes used in the phylogenetic analyses.

Supplementary File 3. Concatenated alignment of amino acid sequences used in the phylogenetic analyses.

Supplementary File 4. Visualization of each metabolic gene cluster detected within the *M. oleifera* genome.